The New England Journal of Statistics in Data Science Volume 3, 242–252 (2025)
DOI: https://doi.org/10.51387/25-NEJSDS95

Bayesian Information Sharing for Equivalence Testing with an Application to Dose Proportionality Studies

Wenru ZHOU, Samantha MAWHINNEY, Peter ANDERSON, and Alexander KAIZER*

Abstract

Dose proportionality is an essential aspect of pharmacokinetics (PK). We aim to enhance the efficiency of PK studies by incorporating interim analyses and utilizing data from past trials to increase precision and enable early termination of studies if applicable. In this paper, we extend the multisource exchangeability model (MEM) to the setting with correlated data with interim analyses. Simulation results indicate that the MEM estimators are efficient even with smaller sample sizes, although smaller sample sizes may have higher mean square error (MSE) and bias due to early stopping and more liberal data borrowing from non-exchangeable supplementary sources. Our recommendation is to use a constrained MEM approach when considering small sample sizes, with additional caution needed around the equivalence boundary to better control the inflated type I error rate, bias, and MSE. This research extends the application of MEMs from linear regression models to settings with correlated data using linear mixed effects regression models. It also innovatively applies MEMs to equivalence testing in the context of dose proportionality studies, thereby enhancing their efficiency.

AMS 2000 subject classifications: Primary 92B15; secondary 62F15.
keywords and phrases: Multisource Exchangeability Model, Interim Monitoring, Dose Proportionality, Equivalence Test.

CONTENTS

1 Introduction . . . 242
2 Background . . . 243
2.1 Dose Proportionality . . . 243
2.2 Equivalence Test . . . 243
2.3 Interim Analysis in Equivalence Test . . . 244
3 Methods . . . 244
3.1 Multisource Exchangeability Models . . . 244
3.2 Group Sequential Design in Bayesian Framework . . . 244
3.3 Linear Mixed Effect Models . . . 245
4 Simulation Design . . . 245
4.1 Correlation Setting . . . 245
4.2 Data Analyses . . . 246
5 Results . . . 247
5.1 Bias and MSE . . . 247
5.2 Power and Type I Error Rates . . . 247
5.3 Averaging Stopping Point . . . 247
5.4 Different Prior Source Exchangeability Weights of the Boundary Case Scenario . . . 248
6 Discussion . . . 248
Supplementary Material . . . 251
Funding . . . 251
References . . . 251
Authors' addresses . . . 252

*Corresponding author.

1. INTRODUCTION

Dose proportionality is a desirable dose-response relationship between dose and pharmacokinetics (PK) response. Dose proportionality indicates that doubling the dose will double the PK response. The common response can be C_{max} or total systemic exposure, also known as, area under the curve (AUC) to the drug. For drugs that are given at a fixed dose and fixed interval, drug concentration at steady state C_{ss} as the PK response is used. This paper is motivated by the real-world application of Tenofovir diphosphate (TFV-DP) in dried blood spots (DBS) based on a 12-week concentration that approaches steady state[2].

There are multiple approaches we could use to improve the efficiency of a PK study. One potential option is to incorporate interim analyses that allow the trial to stop early for futility or efficacy with respect to declaring dose equivalence. However, naive multiple testing will lead to inflated type I error rate[3]. Ad hoc rules aim to preserve the operating characteristics of a study, such as power and type I error rates, even when interim analyses are introduced into the research process [12]. Group sequential tests, introduced by [23], [21, 28, 7, 15], have been wildly applied in the design of clinical studies. These methods ensure that key operating characteristics are retained while enabling multiple data reviews. Such reviews can lead to an early termination of a study if it appears futile (indicating no detectable effect), or if a significant difference is found (demonstrating superiority or inferiority), or both.

If past PK studies with the same drug is available, one may wish to use methods to incorporate past study data into the current study to potentially terminate early if interim analyses are included or to increase the precision of a final estimator. Numerous methodologies target the incorporation of such historical data. These methods can be broadly categorized into two distinct groups based on their mechanism. The first group, including Bayesian hierarchical classification and information sharing[4], calibrated Bayesian hierarchical model[6], and Bayesian cluster hierarchical model[5], seek to foster information borrowing between multiple ongoing studies. The second group, comprising methodologies such as the multisource exchangeability model[16], robust-meta-analytic-predictive prior[25], and power prior[14], are motivated to leverage completed historical data for the benefit of current studies, but may also leverage ongoing studies as well. A comparative analysis of these various methodologies was conducted by [27], further enriching our understanding of these diverse strategies and their potential applications in PK studies.

In this paper we consider the setting where we have prior trial data available. Specifically, we extend the multisource exchangeability model (MEM) developed by [16] and implemented by [19]. Details of the general structure of MEMs and our extensions for their use in equivalence hypothesis testing are described in section 3. In section 4, simulation studies are used to assess the performance of MEMs with interim monitoring for dose proportionality. The outcome is the concentration of drug at steady state, and various dose levels are considered based on a motivating real-world context with two completed PK studies [29, 2]. In our simulations, we assume that there may be one or two historical data sources, which might be exchangeable, "partially" exchangeable, or not exchangeable to the current study. We include interim monitoring as the primary study continues and use MEMs to borrow the information from completed historical trial(s). Results are shown in section 5. Conclusions and future work are described in the final section.

2. BACKGROUND

2.1 Dose Proportionality

The therapeutic window is a range of dose concentrations that will provide an effective response without significant adverse effects. Therefore, studies are designed to carefully determine the dose given to patients so that the concentration of the drug in the human body will fall into the therapeutic window. If the dose is too low, the drug might not be effective. However, if the dose is too high, adverse events might harm the human body.

Therefore, it is essential to precisely predict the drug concentration based on a given dose. Generally, there are three possible relationships between drug concentration and drug dose: (1) Increasing doses will result in a proportional increase in drug concentration. (2) Increasing doses will result in more than proportional increase in drug concentration. (3) Increasing doses will result in less than proportional increase in drug concentration. Among three relationships, (1) represents when dose proportionality is achieved. If dose proportionality is achieved, predicting the drug concentration will be much easier for future applications.

There are many methods to model the relationship between concentration and dose. The power model [13] is one of the most wildly used. It assumes that the logarithm of the pharmacokinetic (PK) parameter is linearly related to logarithm of dose. The PK parameter can be C_{max} or AUC for single dose, or $C_{steadystate}$ for multiple doses. The formula can be written as $ln(PK) = \beta_0 + \beta_1 ln(dose)$. The equation can also be rewritten without the natural log as $PK = e^{\beta_0} dose^{\beta_1}$.

Assuming the maximum dose is h, and the minimum dose is l, $\frac{PK_h}{PK_l} = \frac{e^{\beta_0} h^{\beta_1}}{e^{\beta_0} l^{\beta_1}} = \frac{h^{\beta_1}}{l^{\beta_1}} = (\frac{h}{l})^{\beta_1} = r^{\beta_1}$. If $\beta_1 = 1$, we can obtain $\frac{PK_h}{PK_l} = \frac{h}{l}$. This means that, if the dose is doubled, the drug concentration is also doubled. In other words, dose proportionality is achieved.

2.2 Equivalence Test

In the traditional comparative research study, hypotheses are often proposed to compare a new therapy to a current therapy (two arms) or a meaningful value (single arm) to detect any difference. Therefore, the null hypothesis is that there is no difference between new therapy and the current therapy, or the meaningful value, and the alternative hypothesis is that there is a difference. However, if we apply this approach to the assessment of dose proportionality, rejecting the null hypothesis $\beta_1 = 1$ is equivalent to claiming alternative hypothesis $\beta_1 \neq 1$ is true.

Therefore, this logic is not appropriate for establishing dose proportionality since we need to prove that $\beta_1 = 1$ to establish dose proportionality. Assuming $\beta_1 \neq 1$, strong evidence is needed to reject this assumption. Therefore, we cannot use the traditional comparative hypothesis.

Instead, equivalence testing is suitable for our context. The simplest and most common way is to use the two one-sided test (TOST) procedure described by [26]. It is defined by two margins (θ_l, θ_h) at the left and right side of the value that indicates equivalence. One then performs two one-sided tests at a desired α level: (1) $H_0 : \theta \leq \theta_l$ versus $H_1 : \theta > \theta_l$, and (2) $H_0 : \theta \geq \theta_h$ versus $H_1 : \theta < \theta_h$. This process is equivalent to calculating a $(1-2\alpha)\times 100\%$ confidence interval for θ and determining if it lies entirely between θ_l and θ_h.

Assessment of dose proportionality is ultimately a kind of equivalence test. We want to test $H_0 : r^{\beta_1 - 1} <= \theta_l$ or $r^{\beta_1 - 1} >= \theta_h$ versus $H_1 : \theta_l < r^{\beta_1 - 1} < \theta_h$. This is equivalent to testing if $1 + \frac{ln\theta_l}{lnr} < \beta_1 < 1 + \frac{ln\theta_h}{lnr}$. The U.S. Food and Drug Administration defined $\theta_l = 0.8$ and $\theta_h = 1/\theta_l = 1.25$ [9]. When the $(1-2\alpha)\times 100\%$ CI with $\alpha = 0.05$ for parameter β_1, which is the slope for dose, falls between $1 - \frac{0.223}{lnr} < \beta_1 < 1 + \frac{0.223}{lnr}$, dose proportionality is established.

2.3 Interim Analysis in Equivalence Test

In k-total analyses, the $\alpha_1, \alpha_2, \ldots, \alpha_k$ are spent across different interim analyses to maintain the total α level. For frequentist methods, the confidence interval for β_1 is established at ith interim analysis by calculating $(1-2\alpha_i)\times 100\%$ confidence interval [8]. For Bayesian methods, the credible interval $(L_{\epsilon_1}, U_{\epsilon_1})$ for β_1 is established at ith interim analysis by calculating $P(\theta < L_{\epsilon_1}) = \epsilon_1$ and $P(\theta > L_{\epsilon_2}) = \epsilon_2$ [11]. We set $\epsilon_1 + \epsilon_2 = 2\alpha_i$. And bioequivalence is established if $(L_{\epsilon_1}, U_{\epsilon_1})$ falls between $1 + \frac{ln\theta_l}{lnr}$ and $1 + \frac{ln\theta_h}{lnr}$. While there are many choices of the credible interval, we consider high density intervals (HDI) for the remainder of this manuscript.

3. METHODS

In this section, we introduce the multisource exchangeability model (MEM) approach to sharing information across multiple data sources. We do so in the context of PK studies, so we assume we are going to conduct a study to assess the dose proportionality of a drug.

For our PK study, we will use the power model and use the C_{ss} as the PK response: $ln(C_{ss,p}) = \beta_{0,p} + \beta_{1,p} ln(dose_p) + \epsilon_p$, where $\epsilon_p \sim N(0, E)$ and p represents our current primary study while E representing the variance of error. Meanwhile, there may be similar completed studies about the dose proportionality of our study drug. One may want to borrow from these historical data sources for our primary study if they suspect the studies are exchangeable. For each historic data source, the power model is similarly defined: $ln(C_{ss,h}) = \beta_{0,h} + \beta_{1,h} ln(dose_h) + \epsilon_h$, where $\epsilon_h \sim N(0, E)$ and h represents each supplementary data source $h = 1, 2, \ldots, H$, where H is the total number of historical data sources.

If $\beta_{1,p} = \beta_{1,h}$, it is concluded that the primary data and the h^{th} supplementary data are exchangeable based on the MEM approach [16]. Since there are H historical data sources, and we wish to determine whether each source is either "exchangeable" or "non-exchangeable" with the primary study, there are a total of $K = 2^H$ possible exchangeability configurations defined as Ω_k, where $k = 1, 2, \ldots, K$.

3.1 Multisource Exchangeability Models

First we briefly review the multisource exchangeability model [16]. Assuming there are two historical data sources (D_1, D_2) and one primary study D_p, where D represents the observed data, there will be 4 assumed exchangeability patterns $(\Omega_1, \Omega_2, \Omega_3, \Omega_4)$ since $H = 2$ and $K = 2^2$. For each exchangeability configuration, there will be a posterior weight estimated for its appropriateness of assuming exchangeability with the current trial, and the final model will be a weighted sum of the models. For two historical data sources, the final model utilized for inference will be $q(\beta|D_p, D_1, D_2) = \sum_{k=1}^{4} \omega_k q(\beta|\Omega_k)$, where $q(\beta|\omega_k)$ is the posterior distribution for our parameter of interest from each configuration, β is the parameter of interest, and ω_k is the posterior weight for the configuration. Each ω_k in the MEM framework is estimated using Bayesian model averaging: $\omega_k = P(\Omega_k|D) = \frac{\pi(\Omega_k)P(D|\Omega_k)}{\sum_{j=1}^{K} \pi(\Omega_j)P(D|\Omega_j)}$. However, since there is no closed-form solution when considering regression models, [19] proposed a BIC to approximate this term so that $\omega_k = P(\Omega_k|D) = \frac{\pi(\Omega_k)exp(-0.5\Delta_k)}{\sum_{j=1}^{K} \pi(\Omega_j)exp(-0.5\Delta_j)}$, where $\Delta_k = BIC_k - min(BIC_1, BIC_2, \ldots BIC_K)$. In practice, we note that any other model selection criterion would also work.

Based on prior work by [16], we define a prior π_e as a common probability for each supplementary data source being exchangeable with the primary data source, which is then used to estimate $\pi(\Omega_k)$, the prior probability of a given configuration of exchangeability. Traditionally, π_e can be obtained by calibrating until desired frequentist trial operating characteristics such as type I error rates or power are achieved. In our study, we utilized the value $\pi_e = 0.05$ from [19], in place of fine tuned calibration of π_e, to reduce the potential influence of supplemental data sources for a more conservative approach based on their reported operating characteristics. It is worth noting that while $\pi_e = 0.05$ may seem overly conservative, [19] reported that it provided a good balance between increased power and bias.

3.2 Group Sequential Design in Bayesian Framework

While type I error rate and power are rooted as operating characteristics in the frequentist approach, Bayesian approaches often summarize these operating characteristics because they are requested per U.S. Food and Drug Administration guidance [10]. In this work, the High-Density Interval (HDI) for our Bayesian credible interval (CrI) is used to determine dose proportionality via equivalence testing. If the CrI is entirely contained within the equivalence limits, the dose is declared proportional. Specifically, we used the HDInterval package in R to estimate the HDI with arguments set to require only a single interval returned (i.e., if the HDI were multimodal the package returns the single interval with the highest density that covers the specified probability density) [20].

While numerous approach exist to group sequential designs, we choose a flat decision boundary similar to a Pocock boundary for all MEM approaches with interim monitoring to represent the same approach as proposed by [19]. This approach is implemented by estimating the Pocock boundaries from the gsDesign package and then estimating credible intervals adjusted based on the corresponding Z statistic from gsDesign for use in interim monitoring [1]. To avoid borrowing too much information from historical data when the primary sample size is small at the early stage of interim analysis, we also consider the constrained MEM proposed by [19] as a more conservative approach to borrowing as compared to the "unconstrained" MEM approach is used.

3.3 Linear Mixed Effect Models

Our proposed method originates from the adaptation of MEM to simple linear regression as proposed by [19], applied in conjunction for parallel studies with observations assumed to be independent. This independence assumption, however, may not hold in the context of dose proportionality studies if a crossover design is employed where subjects are observed multiple times across different doses. Consequently, to account for the intra-subject correlations among observations, a linear mixed model (LMM) would be more appropriate. In this section, we propose an extension of MEM to LMMs, specifically tailored for crossover study applications.

In the primary and supplementary datasets considered in our methods, we assume every participant receives two doses, separated by a washout period following the initial dose assignment under the presumption of no carry-over effects. Let $Y_{ij} = ln(C_{ss,ij})$ denote the logarithmic transformation of $C_{ss,ij}$, where i indexes the participants from 1 through n, and j represents the first and second observations. Assume that the jth observation for subject i, Y_{ij}, is distributed as $N(\beta_0 + \beta_1, \sigma_u^2 + \sigma_e^2)$. The covariance between the two observations of the same subject i, $Cov(Y_{i1}, Y_{i2})$, is assumed to be σ_u^2, and the covariance between observations of distinct subjects i and i', $Cov(Y_{ij}, Y_{i'j})$, is assumed to be 0.

For illustration, we utilize a single supplementary resource, bearing in mind that this approach can readily be extended to encompass multiple supplementary resources. The linear mixed model for the non-exchangeability configuration (i.e., assuming the supplemental study is not exchangeable with the current study) is:

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 ln(Dose) + \beta_2 I(S = 1) \\ + \beta_3 ln(Dose) * I(S = 1) + u_i + \epsilon_{ij} \end{aligned} \tag{3.1}$$

where S = 1 means the data is from the supplementary data. For the exchangeability configuration (i.e., assuming the supplemental source is exchangeable with the current study) is:

$$Y_{ij} = \beta_0 + \beta_1 ln(Dose) + \beta_2 I(S = 1) + u_i + \epsilon_{ij} \tag{3.2}$$

where $i = 1, 2, \ldots, N$; $j = 1, 2$; $u_i \sim N(0, \sigma_u^2)$; $\epsilon_{ij} \sim N(0, \sigma_e^2)$; σ_u^2 and σ_e^2 represent the variance of error introduced by different participants and observations, respectively;

To leverage the BIC approximation noted earlier, a set of models are fit using a frequentist approach which maximizes the restricted log-likelihood and then by the Bayesian approach using MCMC. From frequentist models, the BIC for the posterior weight calculation is easy to calculate. From the Bayesian models using MCMC, we have the following process to estimate $\vec{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$:

1. Estimate the posterior distributions of $\vec{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ from each exchangeable pattern. For example, with 1 supplementary source, there are 2 exchangeability patterns (Ω_1, Ω_2) each with its own separate regression model, and there will be 2 sets of $\vec{\beta}$ from 2 different posterior distributions.
2. The final set of $\vec{\beta}$ is sampled from the 2 sets of β based on the weight (ω_1, ω_2) from each model. In other words, the final $\vec{\beta}$ is a mixture distribution from each configuration of exchangeability.

4. SIMULATION DESIGN

To evaluate the performance of MEMs to facilitate information sharing for the primary trial analysis for dose proportionality from completed trials, we conducted a set of simulation studies. For our simulation, we assume $\alpha = 0.05$ (equivalent to two one-sided tests with $\alpha = 0.025$) and confidence intervals or credible intervals of $95\%CI$ since our aim is to compare different frequentist and Bayesian methods. This alpha level is lower than the requirement from FDA. The 95% CI of β is wider than the 90% CI of β, and this change should increase the type I error rate in three methods. However, our aim is to compare three methods, and an overall change among three methods should not impact the comparison results. In this simulation, participants are assigned dose levels of $(25, 50, 75, 100)$. The intent is to ensure that the minimum and maximum doses utilized are relevant and meaningful within the context of dose proportionality studies. For instance, in the study conducted by [29], the minimum and maximum doses reported are 26 fmol/punch and 98 fmol/punch, respectively. Similarly, [2] utilized doses ranging from 27 fmol/punch to 97 fmol/punch in their study. Thus, the dose levels assigned in our simulation study are motivated by a real-world range, thereby maintaining validity in the context of dose proportionality investigations. In the primary study, the dose level before and after the washout period for each participant are randomly selected from four doses. In the supplementary trial, the situation is simplified. Each participant can be given dose level 25 or 75 before the washout period. If the first dose is 25, the dose after the washout period will be 50. If the first dose is 75, the dose after the washout period will be 100.

4.1 Correlation Setting

The outcomes are simulated from a log-normal distribution. We assume that data come from a cross-over study with no carry-over effects based on the design of the real-world trials motivating this work. For the primary study, each participant will be assigned multiple doses at a fixed interval until their plateau is reached. After a washout period, each participant will be given a new dose at a fixed interval until the plateau is reached again. However, not all of the participants will have two records, since if a study terminates early, some participants will only have one period

Table 1. Simulation settings.

Past Trials	Setting	π_e	Primary β_1	Trial 1 $\beta_{3,S1}$	Trial 2 $\beta_{3,S2}$	Sample Size*
1	Nonproportionality	0.05	0.84	0.84		(36, 48)
1	Nonproportionality	0.05	0.75	1		(36, 48)
1	Nonproportionality	0.05, 0.01, 0.001	0.84	1		(36, 48)
1	Proportionality	0.05	0.95	1		(36, 48)
1	Proportionality	0.05, 0.01, 0.001	1	1		(36, 48)
1	Proportionality	0.05	1.05	1		(36, 48)
1	Nonproportionality	0.05	1.16	1		(36, 48)
1	Nonproportionality	0.05	1.25	1		(36, 48)
1	Nonproportionality	0.05	1.5	1		(36, 48)
1	Nonproportionality	0.05	0.84	0.84		(100, 200)
1	Nonproportionality	0.05	0.75	1		(100, 200)
1	Nonproportionality	0.05, 0.01, 0.001	0.84	1		(100, 200)
1	Proportionality	0.05	0.95	1		(100, 200)
1	Proportionality	0.05, 0.01, 0.001	1	1		(100, 200)
1	Proportionality	0.05	1.05	1		(100, 200)
1	Nonproportionality	0.05	1.16	1		(100, 200)
1	Nonproportionality	0.05	1.25	1		(100, 200)
1	Nonproportionality	0.05	1.5	1		(100, 200)
2	Proportionality	0.05	1	1	1	(36,36,48)
2	Proportionality	0.05	1	1.16	1	(36,36,48)
2	Nonproportionality	0.05	1.16	1	1	(36,36,48)
2	Nonproportionality	0.05	1.16	1.16	1	(36,36,48)

*The order for sample size is (primary, supplementary) or (primary, supplementary 1, supplementary 2)

of data. For supplementary data sources, all participants are treated with two doses, with a washout period after the first allocated dose assignment, again assuming no carry-over effects.

In data simulation, the standard error for all data sources is 0.15, which means that the unexplained variance between observations after controlling for dose and data sources is 0.0225. The standard error for observations from the same subject is 0.15, which means that the unexplained variance due to the difference between subjects after controlling for dose and data sources is 0.0225. The intercept is set at $\beta_0 = 0$ since it is not used to evaluate dose proportionality.

All considered scenarios are summarized in Table 1. We set the number of participants in the primary study as 36 and in the supplementary study as 48 to match the sample size in TFB-DP Data [29, 2]. Although large sample sizes are rare in phase 1 study, we still consider scenarios where the number of participants in the primary study as 100 and in the supplementary study as 200 to assess dose proportionality. Since dose level ranges from 25 to 100 and $r = max/min = 4$, $1 + ln(0.8)/ln(4) = 0.84$ and $1 + ln(1.25)/ln(4) = 1.16$. Those two values represent the boundary for defining proportionality, where our CI or CrI must be wholly contained between 0.84 and 1.16 to declare a proportional dose.

For the interim analysis, we only allow stopping for declaring dose proportionality. In other words, if the interim data show a CI or CrI within the (0.84, 1.16) interval, the study will be stopped early and dose proportionality declared. However, if any part of the CI or CrI falls outside of this interval, the study will continue enrolling until the next planned interim analysis.

Details of simulation settings are in the appendices. We also consider the simpler context where there are more traditional parallel arm PK dose studies instead of cross-over studies. The details are included in the appendices.

4.2 Data Analyses

Three sets of models are applied across the different simulation scenarios. For correlated data, a mixed effect model takes the correlation into consideration (Full) and a linear regression ignores the correlation (IC). For uncorrelated data, linear regression is performed (NC). Three approaches are implemented for the simulation studies: Unconstrained MEM (MEM-U), Constrained MEM (MEM-C), and a traditional frequentist approach using a Pocock group sequential boundary (Pocock). We will mainly discuss the results from Full model, and the results from IC and NC will be discussed in the supplementary materials.

For the frequentist approach, REML is applied to obtain the estimated coefficients. For all Bayesian approaches, we set vague priors as $\frac{1}{\sigma_u^2} \sim Gamma(0.001, 0.001)$ and $\frac{1}{\sigma_e^2} \sim Gamma(0.001, 0.001)$. We also set the vague independent normal priors $\beta \sim N(0, 100^2)$ on each regression coefficient. For the selection of π_e we set the probability of

any source being exchangeable equal to 0.05. As a sensitivity analysis, for select scenarios we also set π_e to 0.01 and 0.001 to evaluate the sensitivity of results to source exchangeability prior specification. A total of 1000 simulations are conducted using R v4.2.0 (Vienna, Austria) [24] and RJags[22]

5. RESULTS

In the following subsections we focus on the results for the scenarios with a single supplementary source for the Constrained MEM (MEM-C), Unconstrained MEM (MEM-U), and group sequential method with Pocock boundary. Figure 1 summarizes the bias, mean square error (MSE), Power/Type I Error Rate, and the average interim analysis stopping point for scenarios where the supplementary source has a slope of 1 with varying primary source slopes between 0.75 and 1.50. Results are stratified by the smaller and larger sample size scenarios.

Results for two supplemental sources, the performance when ignoring correlation (IC), the performance with no correlation (NC), and additional figures for the one supplemental source results are presented in the Supplementary Materials.

The Supplementary Materials includes additional tabular and graphical summaries of results and other scenarios.

5.1 Bias and MSE

In Figure 1a, the bias of the Pocock approach is approximately 0 across all primary slopes for both smaller and larger sample sizes because no information is incorporated from the one supplementary source. The bias for MEM-C and MEM-U are approximately 0 either when the primary slope is equal to 1 and is exchangeable with the supplemental source or if the primary slope is equal to 1.5 and is different enough to be decidedly non-exchangeable. For both smaller and larger sample sizes, the most extreme biases of approximately ± 0.025 for MEM-U and ± 0.02 for MEM-C are observed for the 0.84, 0.95, 1.05, and 1.16 primary slope scenarios, pulling the primary slope estimate towards the supplemental source slope of 1. The bias attenuates towards 0 as the primary slope further departs from the dose proportionality slope of 1.

In Figure 1b, the MSE for the Pocock approach is approximately 0.002 across all primary slopes for the smaller sample size scenario and is approximately 0.001 for the larger sample size scenario. When the primary slope is 1, the MSE is approximately 30% lower for both MEM-C and MEM-U for the smaller sample size scenario and is 17% lower for MEM-C and 50% lower for MEM-U in the larger sample size scenario. When the primary source is exchangeable, this represents an area of improved efficiency.

The MSE is larger than the Pocock approach for almost all other scenarios due to the bias introduced when incorporating the non-exchangeable supplementary source, peaking at approximately 2 [3] times higher on the equivalence boundary for MEM-C and 2.75 [6] times for MEM-U in the smaller [larger] sample size scenario. However, when the primary slope is sufficiently different from the primary source at 1.5, MEM-C and MEM-U have the same MSE as the Pocock approach.

5.2 Power and Type I Error Rates

The probability of rejection to non-proportionality is the power under the scenarios of dose proportionality and is the type I error rate under the scenarios of non-proportionality. Figure 1c visualizes these results across different scenarios. When the primary slope is 1 for the smaller sample size scenario with 72 in the primary study and 96 in the supplemental study, the power for the Pocock approach is 78.6% compared to 98.0% for MEM-C and 98.2% for MEM-U. This increase in power of 20% with information sharing shows a potential strength of sharing information. In the larger sample size scenario with 200 and 400 in the primary and supplementary studies, respectively, all methods have 100% power.

In the small sample size scenarios where the primary source has a slope of 0.95 or 1.05, representing doses that may still be considered fairly proportional, the Pocock approach only has up to 56.4% power, compared to 88.0% for MEM-C and 89.0% for MEM-U. For the larger sample sizes, the MEM approaches have over 99% power compared to up to 97.5% for Pocock. These results indicate that in scenarios where you would still likely conclude a nearly proportional dose, information sharing improves power.

In contrast, information sharing does present a risk for inflated type I error rates. On the equivalence boundary of 0.84 or 1.16, Pocock's type I error rate is up to 1.4% compared to 12.3% for MEM-C and 18.1% for MEM-U in the smaller sample size scenario. In the larger sample size scenario, Pocock's type I error rate is up to 1.8% compared to 14.6% for MEM-C and 25.7% for MEM-U. However, the MEM's type I error rates quickly decrease to more acceptable levels of up to 0.3% for MEM-C and 2.1% for MEM-U for primary slopes of 0.75 or 1.25, compare to 0% for Pocock in the small sample size scenario with similar results for the large sample size scenario.

5.3 Averaging Stopping Point

In Figure 1d, the average stopping point for the simulation scenarios is depicted, where a value of 1/2/3/4 represents stopping at 25/50/75/100% of the expected trial enrollment. The only early stopping allowed was for early determination of dose proportionality.

In the smaller [larger] sample size scenario, for the (0.95, 1.0, 1.05) primary source slopes, the smallest average stopping rate was achieved for MEM-U at (2.7, 2.5, 2.7) [(1.5, 1.3, 1.5)], followed by MEM-C at (3.0, 2.8, 3.0) [(1.9, 1.7, 1.9)] and Pocock at (3.8, 3.6, 3.8) [(2.7, 2.4, 2.7)]. This represents a reduction of up to 31% [46%] for MEM-U and 22%

[30%] for MEM-C in the smaller [larger] sample size scenarios relative to the Pocock approach which does not include information sharing.

In the smaller [larger] sample size scenario, for the (1.16, 1.25, 1.50) primary source slopes that represent non-proportionality are largest for the Pocock approach at approximately 4 for both smaller and larger sample size scenarios, indicating the method rarely stops early to incorrectly declare dose proportionality. In contrast, the average stopping point for MEM-U is (3.7, 3.9, 4.0) [(3.3, 3.9, 4.0)] and for MEM-C it is (3.9, 4.0, 4.0) [(3.7, 4.0, 4.0)]. These results suggest early stopping is more likely than for Pocock for the 1.16 and 1.25 slopes. Further, for larger sample sizes, the 1.16 boundary shows an even lower estimate for MEM-U and MEM-C compared to the smaller sample size scenario, suggesting that more information sharing may be present when the supplemental source is twice as large at 400 versus 200 (compared to 96 versus 72 for the smaller sample size scenario).

5.4 Different Prior Source Exchangeability Weights of the Boundary Case Scenario

Given that the poor performance of MEM-C and MEM-U for the scenario when the primary slope is at the boundary of 0.84 and the supplementary slope is at 1 with $\pi_e = 0.05$, we further investigated the behavior of MEMs under this scenario at different prior source exchangeability weights of $\pi_e = 0.01$ and $\pi_e = 0.001$ to see if we can decrease the bias, MSE, and type I error rate at the boundary while still maintaining the power. Figure 2 presents how operating characteristics changed as the weights decreased.

As the weight for primary data increases from 0.95 to 0.99 and to 0.999 (i.e., π_e decreases from 0.05 to 0.001), the bias and MSE all decreased, and the average stopping points are increased. The type I error rate decreased to 0.05 for the small sample size, while power is still maintained above 0.9. However, the type I error rate is still elevated around 0.08 in the large sample size scenario for MEMs. MEM-C has generally better balanced performance than MEM-U since it has a lower bias, MSE, and type I error rate, while still having reasonable power. In addition, as the weight for primary data goes up, the lines from the MEM-U converges to MEM-C.

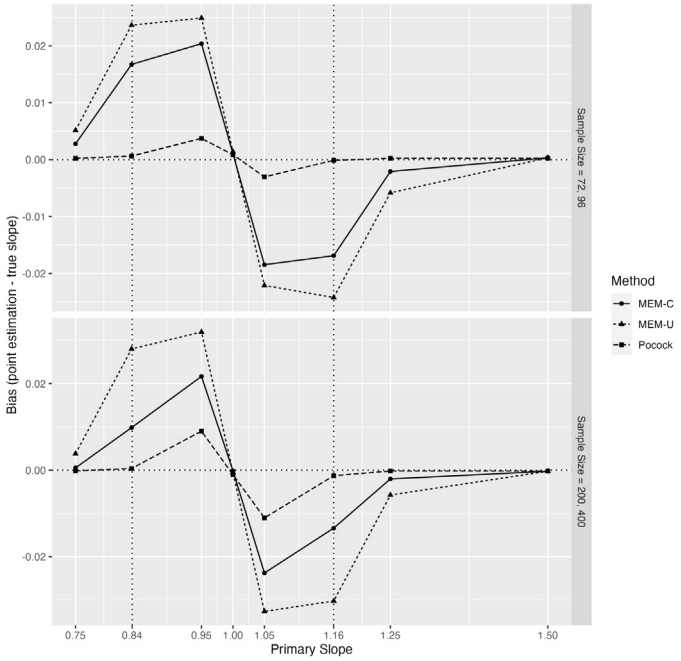
6. DISCUSSION

The proposed research is innovation in multiple ways. First, it extends MEMs from previous work by [19] from linear regression models to a setting with correlated data utilizing linear mixed effects regression models. Second, this is a unique application of MEMs to evaluate hypotheses of equivalence, whereas previous work has been in more traditional superiority testing settings. Finally, it applies these methods to the context of dose proportionality studies to address a novel way of improving their efficiency with interim monitoring and information sharing.

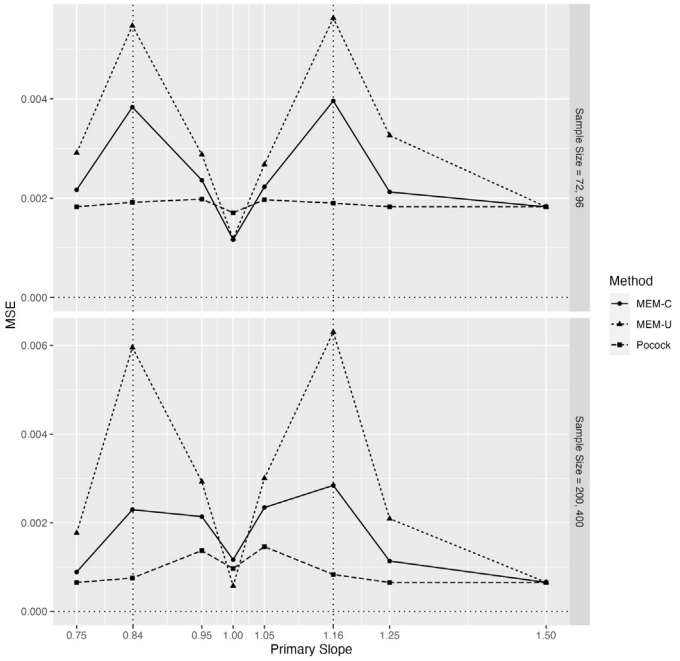
In our paper, we identify that MEM estimators can be more efficient than an approach that does not include information sharing, even when the sample size is small. However, this increased efficiency is at the expense of higher MSE and bias as the primary source departs from the truly dose proportional slope of 1. The use of interim analyses provides another option for improved efficiency in the design of PK studies, especially since incorporating supplemental information may provide evidence that a study could terminate early because equivalence may be shown prior to reaching the maximum sample size. However, in some cases interim monitoring may be challenging due to long periods of follow-up for outcome collection or result in potentially less precision than a trial which did not stop early. For that reason, we note that interim monitoring, while a novel aspect of the proposed design, is not a requirement and one could implement a design without interim monitoring that incorporates historic information from past trials at the end of the current trial. In practice, one must still consider the risk when the primary slope is at the equivalence boundary and at least one supplementary slope is considered dose proportional but is not exchangeable, because MEM approaches borrowed more results from supplementary data sources, especially the more liberal unconstrained MEM. The latter reason can also lead to inflated type I error rates. While utilizing past trials with the same drug and same population are ideal for information sharing, it may be possible to use related drugs if similar PK profiles have been shown or the same drug in different populations, but caution must be taken and ultimately consultation with an expert in the drug class and PK results considered.

Our recommendation is that one should use the constrained MEM when the sample size is small, however, additional steps should be made to avoid the instability around the boundary to control the inflated type I error rate, bias, and MSE in case the primary slope is at boundary and supplementary slopes are not exchangeable. For example, at each interim analysis, before using MEMs to borrow information, we could calculate the point estimate from the primary data itself to check the point estimate and if it is near our equivalence boundary, approaches that do not borrow data may be desired instead. Presumably, any supplemental studies would have either identified dose proportionality or had encouraging results to warrant a subsequent study.

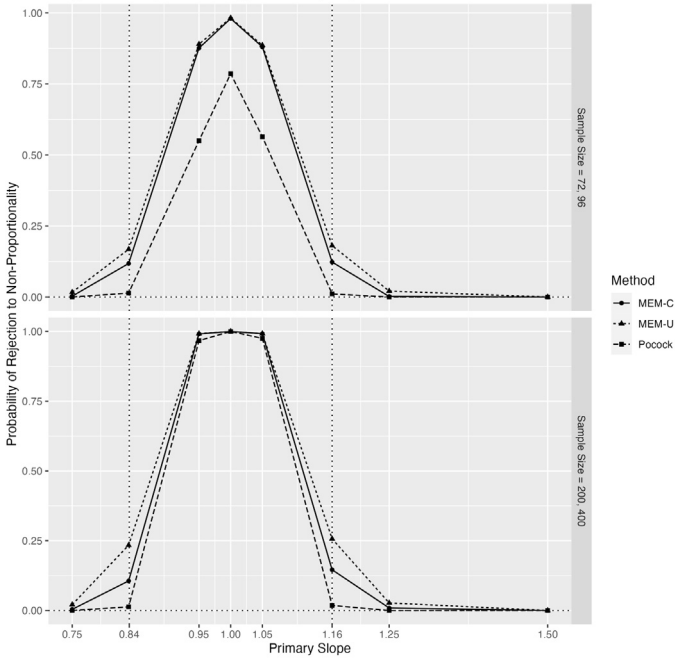
One limitation of this work is that we applied Pocock-style boundaries for both frequentist and Bayesian approaches without further calibration. This may have impacted the results and resulted in higher type I error rates. Fortunately, many options could be considered for further calibration. One approach that we examined for the boundary simulation scenarios was to reduce π_e to 0.001, which helped to reduce the type I error rate, MSE, and bias. As for the threshold of adjustment for the constrained MEM, we used $n_{P,i+1}$ as the threshold of adjustment, however, it can be changed and it may be more efficient than changing



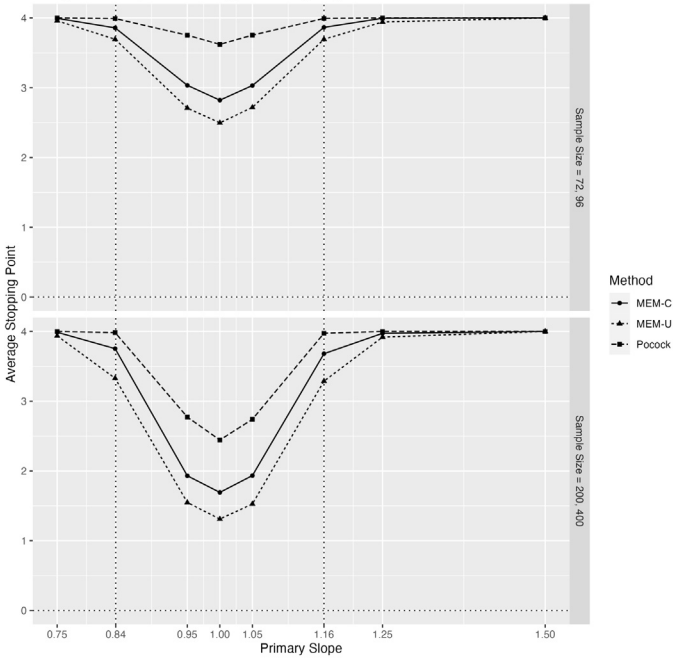
(a) Bias



(b) MSE

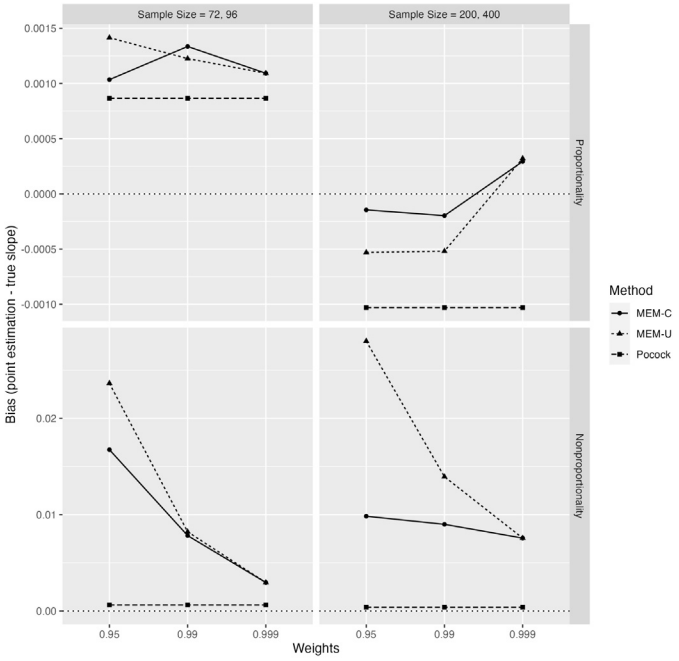


(c) Power/Type I Error Rate

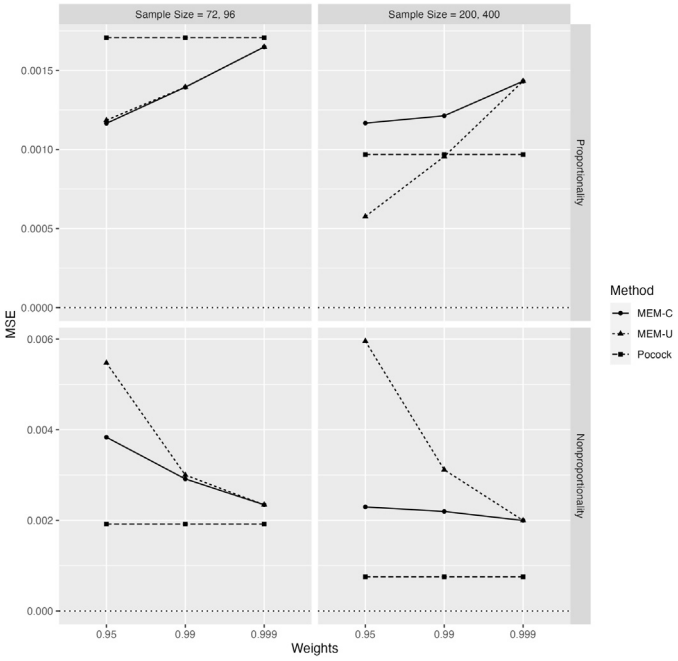


(d) Average Stopping Point

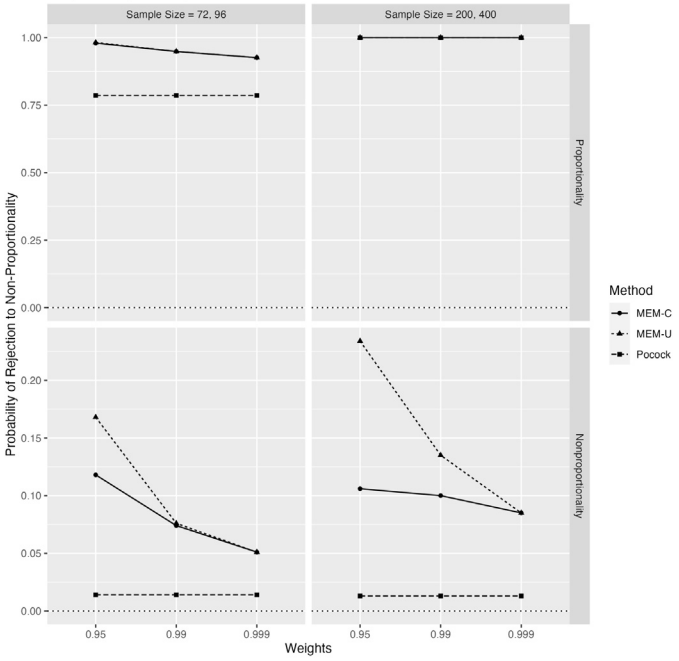
Figure 1: π_e is set to be 0.05 so that the prior weight for the non-exchangeable model is 0.95. Supplementary slope is 1. Different primary slopes are on x-axis. Primary slope 0.75, 0.84, 1.16, 1.25, and 1.5 are for type I error check. Primary slope 0.95, 1, and 1.05 are for power check. Important statistics are on the y-axis. The solid line is for Constrained MEM, short-dashed line is for unconstrained MEM, and long dashed line is for Pocock.



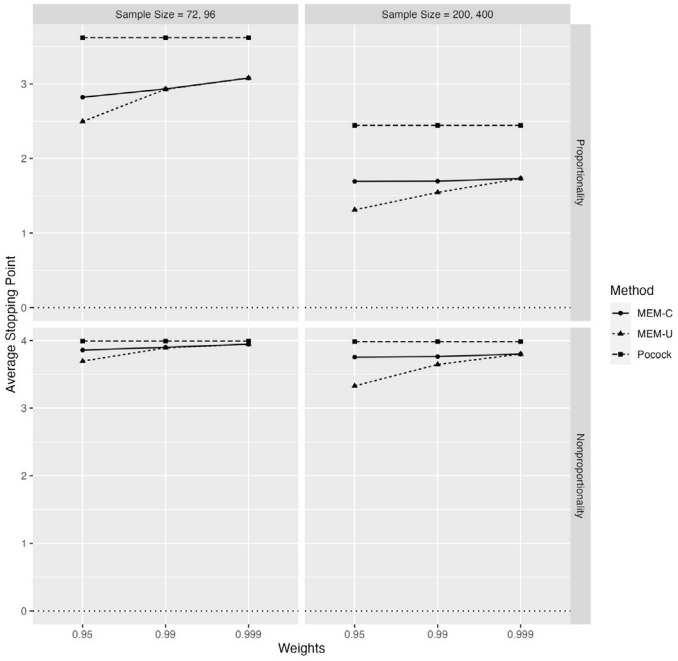
(a) Bias



(b) MSE



(c) Power/Type I Error Rate



(d) Average Stopping Point

Figure 2: Slope from primary data is 1 for power check, and 0.84 for type I error check at boundary, and slope for supplementary data is 1 to check the strength of borrowing of MEM when the primary and supplementary data are not exchangeable. Different prior weights 0.95, 0.99, and 0.999 are put for non-exchangable model at x-axis, which corresponding to $\pi_e = 0.05, 0.01, 0.001$. Important statistics are on the y-axis. Solid line is for Constrained MEM, short-dashed line is for unconstrained MEM, and long dashed line is for Pocock.

π_e since this will limit the sharing of information. For example, we could change the threshold to $n_{P,i+1}/2$. Additionally, one could modify the stopping boundary from Pocock-style to O'Brien-Fleming or other shaped boundaries.

If one is interested in calibrating π_e, we recommend extensive simulation studies to identify where desired frequentist operating characteristics may be achieved. For example, while we used $\pi_e = 0.05$ based on the prior work by [19], work by [17] in the context of basket trials identified acceptable performance with a prior analogous to $\pi_e = 0.10$. Given the uncertainty of selecting a single scenario for hyperparameter calibration, [18] proposed a flexible approach to average across multiple proposed scenarios and in their example also identified the "optimal" prior to be analogous to $\pi_e = 0.1$ to induce information sharing to improve statistical power while minimizing bias and inflated type I error rates.

Another challenge in the proposed design is the potential for carry-over effects between the first and second period of the cross-over trial. While we assume no carry-over effects based on the lack of observed carry-over in the motivating real-world study, this assumption may be overly strong for other studies. In practice, one should carefully design any cross-over trial to ensure a sufficient washout period to avoid carry-over effects. If carry-over effects are observed, there are statistical approaches that may better account for the potential effect between the two periods. In situations with large uncertainty regarding the length of the washout period, a more traditional design without a cross-over element may be more appropriate.

Looking forward, we plan to expand the outcome parameters from the current continuous model to include binary, count, and time-to-event models. Given that the linear mixed model seems to favor the exchangeable model over the linear model, it will be intriguing to see how this preference manifests in the generalized linear model. Additionally, there are opportunities to broaden the multi-source exchangeability model to handle multiple outcomes from joint distributions and to identify efficient approaches to calibrate the various model parameters. We also believe that future work that incorporates a borrow/no borrow decision if the primary study falls into certain challenging cases, such as our simulation scenarios where the primary study fell exactly on the equivalence boundary, may be beneficial. In this case, one may wish to avoid incorporating any supplemental information because the evidence in the primary study is likely to bias our results towards dose proportionality (i.e., increased type I error rates) when we truly fall on the boundary. While one may also choose a more conservative prior to limit borrowing, this will in turn limit the potential gains in efficiency when both primary and supplemental trials are exchangeable with proportional doses, suggesting an interim step to not allow borrowing in certain ranges of point estimates may be more flexible.

SUPPLEMENTARY MATERIAL

Supplementary Materials for Bayesian Information Sharing and Interim Efficacy Monitoring for Equivalence Testing with an Application to Dose Proportionality Studies

FUNDING

AMK and WZ supported by NHLBI K01 HL151754

Accepted 19 October 2025

REFERENCES

[1] Anderson, K. (2024). gsDesign: Group Sequential Design. R package version 3.6.2. https://CRAN.R-project.org/package=gsDesign.

[2] Anderson, P. L., Liu, A. Y., Castillo-Mancilla, J. R., Gardner, E. M., Seifert, S. M., McHugh, C., Wagner, T., Campbell, K., Morrow, M., Ibrahim, M. et al. (2018). Intracellular tenofovir-diphosphate and emtricitabine-triphosphate in dried blood spots following directly observed therapy. *Antimicrobial agents and chemotherapy* **62**(1) 01710-17.

[3] Armitage, P., McPherson, C. and Rowe, B. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)* **132**(2) 235–244. https://doi.org/10.2307/2343787. MR0250405

[4] Chen, N. and Lee, J. J. (2019). Bayesian hierarchical classification and information sharing for clinical trials with subgroups and binary outcomes. *Biometrical Journal* **61**(5) 1219–1231. https://doi.org/10.1002/bimj.201700275. MR4013344

[5] Chen, N. and Lee, J. J. (2020). Bayesian cluster hierarchical model for subgroup borrowing in the design and analysis of basket trials with binary endpoints. *Statistical Methods in Medical Research* **29**(9) 2717–2732. https://doi.org/10.1177/0962280220910186. MR4129440

[6] Chu, Y. and Yuan, Y. (2018). A Bayesian basket trial design using a calibrated Bayesian hierarchical model. *Clinical Trials* **15**(2) 149–158.

[7] Demets, D. L. and Lan, K. G. (1994). Interim analysis: the alpha spending function approach. *Statistics in medicine* **13**(13-14) 1341–1352.

[8] Durrleman, S. and Simon, R. (1990). Planning and monitoring of equivalence studies. *Biometrics* 329–336. https://doi.org/10.2307/2531438. MR1060592

[9] Food and Drug Administration (2001). Statistical approaches to establishing bioequivalence. *Center for Drug Evaluation and Research*.

[10] Food and Drug Administration (2010). Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials.

[11] Freedman, L. S. and Spiegelhalter, D. J. (1989). Comparison of Bayesian with group sequential methods for monitoring clinical trials. *Controlled clinical trials* **10**(4) 357–367.

[12] Friedman, L. M., Furberg, C. D., DeMets, D. L., Reboussin, D. M. and Granger, C. B. (2015) *Fundamentals of clinical trials*. Springer.

[13] Gough, K., Hutchison, M., Keene, O., Byrom, B., Ellis, S., Lacey, L. and McKellar, J. (1995). Assessment of dose proportionality: report from the statisticians in the pharmaceutical industry/pharmacokinetics UK joint working party. *Drug Information Journal* **29**(3) 1039–1048.

[14] Hobbs, B. P., Carlin, B. P., Mandrekar, S. J. and Sargent, D. J. (2011). Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics* **67**(3) 1047–1056. https://doi.org/10.1111/j.1541-0420.2011.01564.x. MR2829239

[15] Jennison, C. and Turnbull, B. W. (1999) *Group sequential methods with applications to clinical trials*. CRC Press. MR1710781

[16] Kaizer, A. M., Koopmeiners, J. S. and Hobbs, B. P. (2018). Bayesian hierarchical modeling based on multisource exchangeability. *Biostatistics* **19**(2) 169–184. https://doi.org/10.1093/biostatistics/kxx031. MR3799610

[17] Kaizer, A. M., Koopmeiners, J. S., Kane, M. J., Roychoudhury, S., Hong, D. S. and Hobbs, B. P. (2019). Basket designs: statistical considerations for oncology trials. *JCO Precision Oncology* **3** 1–9.

[18] Kaizer, A. M., Koopmeiners, J. S., Chen, N. and Hobbs, B. P. (2021). Statistical design considerations for trials that study multiple indications. *Statistical methods in medical research* **30**(3) 785–798. https://doi.org/10.1177/0962280220975187. MR4236836

[19] Kotalik, A., Vock, D. M., Hobbs, B. P. and Koopmeiners, J. S. (2022). A group-sequential randomized trial design utilizing supplemental trial data. *Statistics in Medicine* **41**(4) 698–718. https://doi.org/10.1002/sim.9249. MR4386975

[20] Meredith, M. and Kruschke, J. (2022). HDInterval: Highest (Posterior) Density Intervals. R package version 0.2.4. https://CRAN.R-project.org/package=HDInterval.

[21] O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* 549–556.

[22] Plummer, M. (2022). rjags: Bayesian Graphical Models using MCMC. R package version 4-13. https://CRAN.R-project.org/package=rjags.

[23] Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**(2) 191–199.

[24] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[25] Schmidli, H., Gsteiger, S., Roychoudhury, S., O'Hagan, A., Spiegelhalter, D. and Neuenschwander, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70**(4) 1023–1032. https://doi.org/10.1111/biom.12242. MR3295763

[26] Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of pharmacokinetics and biopharmaceutics* **15** 657–680.

[27] Su, L., Chen, X., Zhang, J. and Yan, F. (2022). Comparative study of bayesian information borrowing methods in oncology clinical trials. *JCO Precision Oncology* **6** 2100394.

[28] Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* 193–199. https://doi.org/10.2307/2531959. MR0882780

[29] Yager, J., Castillo-Mancilla, J., Ibrahim, M. E., Brooks, K. M., McHugh, C., Morrow, M., McCallister, S., Bushman, L. R., MaWhinney, S., Kiser, J. J. et al. (2020). Intracellular tenofovir-diphosphate and emtricitabine-triphosphate in dried blood spots following tenofovir alafenamide: the TAF-DBS study. *JAIDS Journal of Acquired Immune Deficiency Syndromes* **84**(3) 323–330.

Wenru Zhou. Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, United States.
E-mail address: wenru.zhou@cuanschutz.edu

Samantha MaWhinney. Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, United States.
E-mail address: sam.mawhinney@cuanschutz.edu

Peter Anderson. Department of Pharmaceutical Sciences, University of Colorado Anschutz Medical Campus, United States.
E-mail address: peter.anderson@cuanschutz.edu

Alexander Kaizer. Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus, United States.
E-mail address: alex.kaizer@cuanschutz.edu